RandProx: Primal–Dual Optimization Algorithms with Randomized Proximal Updates

[ICLR 2023]

Laurent Condat

Peter Richtárik

King Abdullah Univ. of Science and Technology (KAUST)
Saudi Arabia

Convex optimization

$$\text{Find } x^\star \in \underset{x\in\mathcal{X}}{\arg\min} \sum_{i=1}^{n} f_i(K_i x)$$

with

- linear operators $K_i : \mathcal{X} \to \mathcal{U}_i$
- finite-dimensional real Hilbert spaces $\mathcal{X}, \mathcal{U}_i$
- convex, proper, lower semicontinuous functions $f_i : \mathcal{U}_i \to \mathbb{R} \cup \{+\infty\}$

Convex optimization

$$\text{Find } x^\star \in \underset{x \in \mathcal{X}}{\arg\min} \sum_{i=1}^{n} f_i(K_i x)$$

☞ use a proximal splitting algorithm, with activation of K_i, K_i^*, the gradient or proximity operator of f_i.

$$\text{prox}_f : x \in \mathcal{X} \mapsto \underset{x' \in \mathcal{X}}{\arg\min} \left(f(x') + \tfrac{1}{2}\|x - x'\|^2 \right)$$

Convex optimization

$$\text{Find } x^\star \in \operatorname*{arg\,min}_{x\in\mathcal{X}} \left(f(x) + g(x) + \sum_{i=1}^{n} h_i(K_i x) \right)$$

with:

- f smooth with L-Lipschitz grad $\rightarrow$ calls to ∇f
- calls to $\mathrm{prox}_{\gamma g}$, $\mathrm{prox}_{\tau h_i}$, K_i, K_i^*

Product space trick

$$\text{Find } x^\star \in \underset{x\in\mathcal{X}}{\arg\min} \Big({\color{blue}f(x)} + {\color{green}g(x)} + {\color{red}h(Kx)} \Big)$$

$$\boldsymbol{h}(\boldsymbol{u}) = \sum_{i=1}^{n} h_i(u_i)$$

$$\boldsymbol{K}x = (K_1x, \dots, K_nx)$$

☞ $$\boldsymbol{h}(\boldsymbol{K}x) = \sum_{i=1}^{n} h_i(K_ix)$$

Minimization of 3 functions

$$\text{Find } x^\star \in \arg\min_{x\in\mathcal{X}} \Big(f(x) + g(x) + h(Kx) \Big)$$

$$\downarrow \quad \downarrow \quad \downarrow$$

$$\nabla f, \quad \mathrm{prox}_{\gamma g}, \ \mathrm{prox}_{\tau h}, \ K, \ K^*$$

Dual problem:

$$\text{Find } u^\star \in \arg\min_{u\in\mathcal{U}} \Big((f+g)^*(-K^* u) + h^*(u) \Big)$$

We suppose that there exists $x^\star \in \mathcal{X}$ such that

$$0 \in \nabla f(x^\star) + \partial g(x^\star) + K^* \partial h(Kx^\star).$$

Minimization of 3 functions

$$\text{Find } x^\star \in \operatorname*{arg\,min}_{x\in\mathcal{X}} \Big(f(x) + g(x) + h(Kx) \Big)$$

⚠ $\text{prox}_{\tau h}$

can be costly

Randomized algorithms

$$\text{Find } x^\star \in \arg\min_{x \in \mathcal{X}} \Big(f(x) + g(x) + h(Kx) \Big)$$

↓

randomize ∇f

☞ SGD-type algorithms

The power of randomness

Find $x^\star \in \arg\min_{x \in \mathcal{X}} f(x) = \sum_{i=1}^{n} f_i(x)$ using the ∇f_i

(every f_i is L-smooth and μ-strongly convex)

☞ lower bounds in Woodworth & Srebro [2016]

- deterministic algorithms: $\Omega\big(n\sqrt{L/\mu}\log \epsilon^{-1}\big)$
- randomized algorithms: $\Omega\big((n + \sqrt{nL/\mu})\log \epsilon^{-1}\big)$

Randomized algorithms

$$\text{Find } x^\star \in \operatorname*{arg\,min}_{x\in\mathcal{X}} \Big(f(x) + g(x) + h(Kx) \Big)$$

↓

randomize $\operatorname{prox}_{\tau h}$

?

Proximal splitting algorithms

minimize $f + g + h \circ K$

	$f + g$	☞	forward-backward alg.
1979	$g + h$	☞	Douglas-Rachford alg. / ADMM
2011	$g + h \circ K$	☞	Chambolle-Pock
	$f + h \circ K$	☞	PAPC
2013	$f + g + h \circ K$	☞	Condat, Vu
2017	$f + g + h$	☞	Davis-Yin
2018	$f + g + h \circ K$	☞	PD3O
2022	$f + g + h \circ K$	☞	PDDY

LC et al. “Proximal Splitting Algorithms for Convex Optimization: A Tour of Recent Advances, with New Twists,” *SIAM Review*, 2023

Proximal splitting algorithms

minimize $f + g + h \circ K$

	$f + g$	☞ forward-backward alg.
1979	$g + h$	☞ Douglas-Rachford alg. / ADMM
2011	$g + h \circ K$	☞ Chambolle-Pock
	$f + h \circ K$	☞ PAPC
2013	$f + g + h \circ K$	☞ Condat, Vu
2017	$f + g + h$	☞ Davis-Yin
2018	$f + g + h \circ K$	☞ PD3O
2022	$f + g + h \circ K$	☞ PDDY / AFBA

Salim, LC et al., "Dualize, split, randomize: Fast nonsmooth optimization algorithms," *JOTA*, 2022

PDDY

PDDY

input: initial points $x^0 \in \mathcal{X}$, $u^0 \in \mathcal{U}$;
stepsizes $\gamma > 0$, $\tau > 0$
for $t = 0, 1, \ldots$ **do**

$$\hat{x}^t := \mathrm{prox}_{\gamma g}\left(x^t - \gamma \nabla f(x^t) - \gamma K^* u^t\right)$$
$$u^{t+1} := \mathrm{prox}_{\tau h^*}\left(u^t + \tau K \hat{x}^t\right)$$
$$x^{t+1} := \hat{x}^t - \gamma K^*(u^{t+1} - u^t)$$

end for

Theorem 2. *If $\gamma \in (0, 2/L)$, $\tau > 0$, $\gamma\tau\|K\|^2 \leq 1$, then $(x^t)_{t\in\mathbb{N}}$ converges to a primal solution $x^\star$ and $(u^t)_{t\in\mathbb{N}}$ converges to a dual solution $u^\star$.*

LC, Malinovsky, Richtárik, "Distributed Proximal Splitting Algorithms with Rates and Acceleration," *Frontiers in Signal Processing*, 2022

PDDY

PDDY

input: initial points $x^0 \in \mathcal{X}$, $u^0 \in \mathcal{U}$;
stepsizes $\gamma > 0$, $\tau > 0$
for $t = 0, 1, \ldots$ **do**

$$\hat{x}^t := \mathrm{prox}_{\gamma g}\big(x^t - \gamma \nabla f(x^t) - \gamma K^* u^t\big)$$

$$u^{t+1} := \mathrm{prox}_{\tau h^*}\big(u^t + \tau K \hat{x}^t\big)$$

$$x^{t+1} := \hat{x}^t - \gamma K^*(u^{t+1} - u^t)$$

end for

⚠ $\mathrm{prox}_{\tau h^*}$ can be costly

RandProx

RandProx

input: initial points $x^0 \in \mathcal{X}$, $u^0 \in \mathcal{U}$;
stepsizes $\gamma > 0$, $\tau > 0$; $\omega \geq 0$
for $t = 0, 1, \dots$ **do**

$\hat{x}^t := \mathrm{prox}_{\gamma g}\big(x^t - \gamma \nabla f(x^t) - \gamma K^* u^t\big)$

$u^{t+1} := u^t + \frac{1}{1+\omega} \mathcal{R}^t\big(\mathrm{prox}_{\tau h^*}(u^t + \tau K \hat{x}^t) - u^t\big)$

$x^{t+1} := \hat{x}^t - \gamma(1+\omega) K^*(u^{t+1} - u^t)$

end for

$$\mathbb{E}\left[\mathcal{R}^t(d^t)\right] = d^t \quad \text{and} \quad \mathbb{E}\left[\left\|\mathcal{R}^t(d^t) - d^t\right\|^2\right] \leq \omega \left\|d^t\right\|^2$$

$\mathcal{R}^t \equiv \mathrm{Id}$, $\omega = 0$ ☞ RandProx = PDDY

Linear convergence

Theorem 1. *Suppose that* $\mu_f > 0$ *or* $\mu_g > 0$, *and* $\mu_{h^*} > 0$. *In RandProx, suppose that* $\gamma \in (0, 2/L)$, $\tau > 0$, $\gamma\tau\big((1-\zeta)\|K\|^2 + \omega_{\mathrm{ran}}\big) \leq 1$. *Then* $\forall t \geq 0$,

$$\mathbb{E}\left[\Psi^t\right] \leq c^t \Psi^0, \text{ where}$$

$$\Psi^t = \frac{1}{\gamma}\left\|x^t - x^\star\right\|^2 + (1+\omega)\left(\frac{1}{\tau} + 2\mu_{h^*}\right)\left\|u^t - u^\star\right\|^2,$$

$$c = \max\left(\frac{(1-\gamma\mu_f)^2}{1+\gamma\mu_g}, \frac{(1-\gamma L)^2}{1+\gamma\mu_g}, 1 - \frac{2\tau\mu_{h^*}}{(1+\omega)(1+2\tau\mu_{h^*})}\right).$$

Moreover, $(x^t)_{t\in\mathbb{N}}$ *converges to* $x^\star$ *and* $(u^t)_{t\in\mathbb{N}}$ *converges to* $u^\star$, *almost surely.*

Linear convergence

Theorem 2. *Suppose that* $g = 0$, $\mu_f > 0$, *and that* $\lambda_{\min}(KK^*) > 0$ *or* $\mu_{h^*} > 0$. *In RandProx, suppose that* $\gamma \in (0, 2/L)$, $\tau > 0$, $\gamma\tau\big((1-\zeta)\|K\|^2 + \omega_{\text{ran}}\big) \leq 1$. *Then* $\forall t \geq 0$,

$$\mathbb{E}\left[\Psi^t\right] \leq c^t \Psi^0, \textit{ where}$$

$$\Psi^t = \frac{1}{\gamma}\left\|x^t - x^\star\right\|^2 + (1+\omega)\left(\frac{1}{\tau} + 2\mu_{h^*}\right)\left\|u^t - u^\star\right\|^2,$$

$$c = \max\left((1-\gamma\mu_f)^2, (1-\gamma L)^2, 1 - \frac{2\tau\mu_{h^*} + \gamma\tau\lambda_{\min}(KK^*)}{(1+\omega)(1+2\tau\mu_{h^*})}\right).$$

Moreover, $(x^t)_{t\in\mathbb{N}}$ *converges to* $x^\star$ *and* $(u^t)_{t\in\mathbb{N}}$ *converges to* $u^\star$, *almost surely.*

Examples

RandProx-skip

input: initial points $x^0 \in \mathcal{X}$, $u^0 \in \mathcal{U}$;
stepsizes $\gamma > 0$, $\tau > 0$; $p \in (0, 1]$
for $t = 0, 1, \ldots$ **do**
  $\hat{x}^t := \mathrm{prox}_{\gamma g}(x^t - \gamma \nabla f(x^t) - \gamma K^* u^t)$
  Flip a coin θ^t = (1 with prob. p, 0 else)
  if $\theta^t = 1$ **then**
    $u^{t+1} := \mathrm{prox}_{\tau h^*}(u^t + \tau K \hat{x}^t)$
    $x^{t+1} := \hat{x}^t - \frac{\gamma}{p} K^*(u^{t+1} - u^t)$
  else
    $u^{t+1} := u^t$, $x^{t+1} := \hat{x}^t$
  end if
end for

$$\mathcal{R}^t : d^t \mapsto \begin{cases} \frac{1}{p} d^t & \text{with prob } p \\ 0 & \text{with prob } 1-p \end{cases}$$

$$\omega = \frac{1}{p} - 1$$

Examples

RandProx-skip

input: initial points $x^0 \in \mathcal{X}$, $u^0 \in \mathcal{U}$;
stepsizes $\gamma > 0$, $\tau > 0$; $p \in (0, 1]$
for $t = 0, 1, \ldots$ **do**
  $\hat{x}^t := \mathrm{prox}_{\gamma g}(x^t - \gamma \nabla f(x^t) - \gamma K^* u^t)$
  Flip a coin θ^t = (1 with prob. p, 0 else)
  if $\theta^t = 1$ **then**
    $u^{t+1} := \mathrm{prox}_{\tau h^*}(u^t + \tau K \hat{x}^t)$
    $x^{t+1} := \hat{x}^t - \frac{\gamma}{p} K^*(u^{t+1} - u^t)$
  else
    $u^{t+1} := u^t$, $x^{t+1} := \hat{x}^t$
  end if
end for

Example: $g = 0$, $\mu_{h^*} = 0$, $K = \mathrm{Id}$, $\tau = \frac{p}{\gamma}$, $\gamma = \frac{1}{L}$

iter. complexity:
$\mathcal{O}\left(\max\left(\frac{L}{\mu}, \frac{1}{p^2}\right) \times \log(\epsilon^{-1})\right)$

prox. complexity:
$\mathcal{O}\left(\max\left(\frac{pL}{\mu}, \frac{1}{p}\right) \times \log(\epsilon^{-1})\right)$

Examples

$\min\ f + g + \sum_{i=1}^n h_i$

RandProx-minibatch

input: initial points $x^0 \in \mathcal{X}$, $(u_i^0)_{i=1}^n \in \mathcal{X}^n$;
stepsize $\gamma > 0$; $k \in \{1, \dots, n\}$
$v^0 := \sum_{i=1}^n u_i^0$
for $t = 0, 1, \dots$ **do**
  $\hat{x}^t := \mathrm{prox}_{\gamma g}\big(x^t - \gamma \nabla f(x^t) - \gamma v^t\big)$
  pick $\Omega^t \subset \{1, \dots, n\}$ of size k unif. at random
  for $i \in \Omega^t$ **do**
    $u_i^{t+1} := \mathrm{prox}_{\frac{1}{\gamma n} h_i^*}(u_i^t + \frac{1}{\gamma n}\hat{x}^t)$
  end for
  for $i \in \{1, \dots, n\} \backslash \Omega^t$ **do**
    $u_i^{t+1} := u_i^t$
  end for
  $v^{t+1} := \sum_{i=1}^n u_i^{t+1}$
  $x^{t+1} := \hat{x}^t - \frac{\gamma n}{k}(v^{t+1} - v^t)$
end for

$\mathcal{R}^t$: sampling

$\omega = \frac{n}{k} - 1$, $\tau = \frac{1}{\gamma n}$

Examples

RandProx-FL

input: initial estimates $(x_i^0)_{i=1}^n \in \mathcal{X}^n$, $(u_i^0)_{i=1}^n \in \mathcal{X}^n$ such that $\sum_{i=1}^n u_i^0 = 0$; stepsize $\gamma > 0$; $\omega \geq 0$
for $t = 0, 1, \ldots$ **do**
  for $i = 1, \ldots, n$ at nodes in parallel **do**
    $\hat{x}_i^t := x_i^t - \gamma \nabla f_i(x_i^t) - \gamma u_i^t$
    $a_i^t := \mathcal{R}^t(\hat{x}_i^t)$
    // send compressed vector a_i^t to master
  end for
  $a^t := \frac{1}{n} \sum_{i=1}^n a_i^t$ // aggregation at master
  // broadcast a^t to all nodes
  for $i = 1, \ldots, n$ at nodes in parallel **do**
    $d_i^t := a_i^t - a^t$
    $u_i^{t+1} := u_i^t + \frac{1}{\gamma(1+\omega)^2} d_i^t$
    $x_i^{t+1} := \hat{x}_i^t - \frac{1}{1+\omega} d_i^t$
  end for
end for

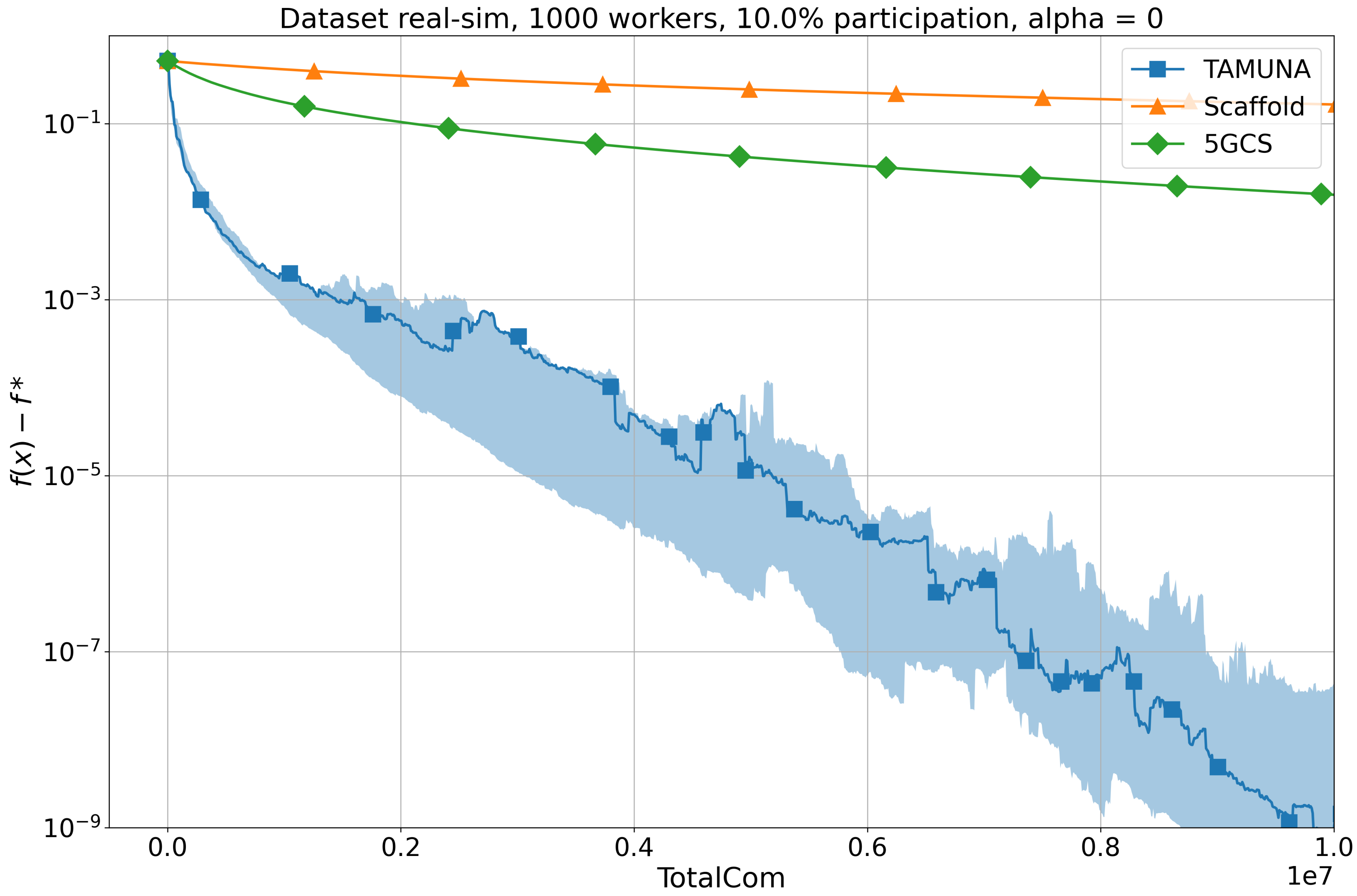
$\min \sum_{i=1}^n f_i$

$\mathcal{R}^t$: linear compression

Extension

Decoupled primal and dual randomization

LC et al. “TAMUNA: Doubly accelerated federated learning with local training, compression, and partial participation,” preprint, 2023



Conclusion

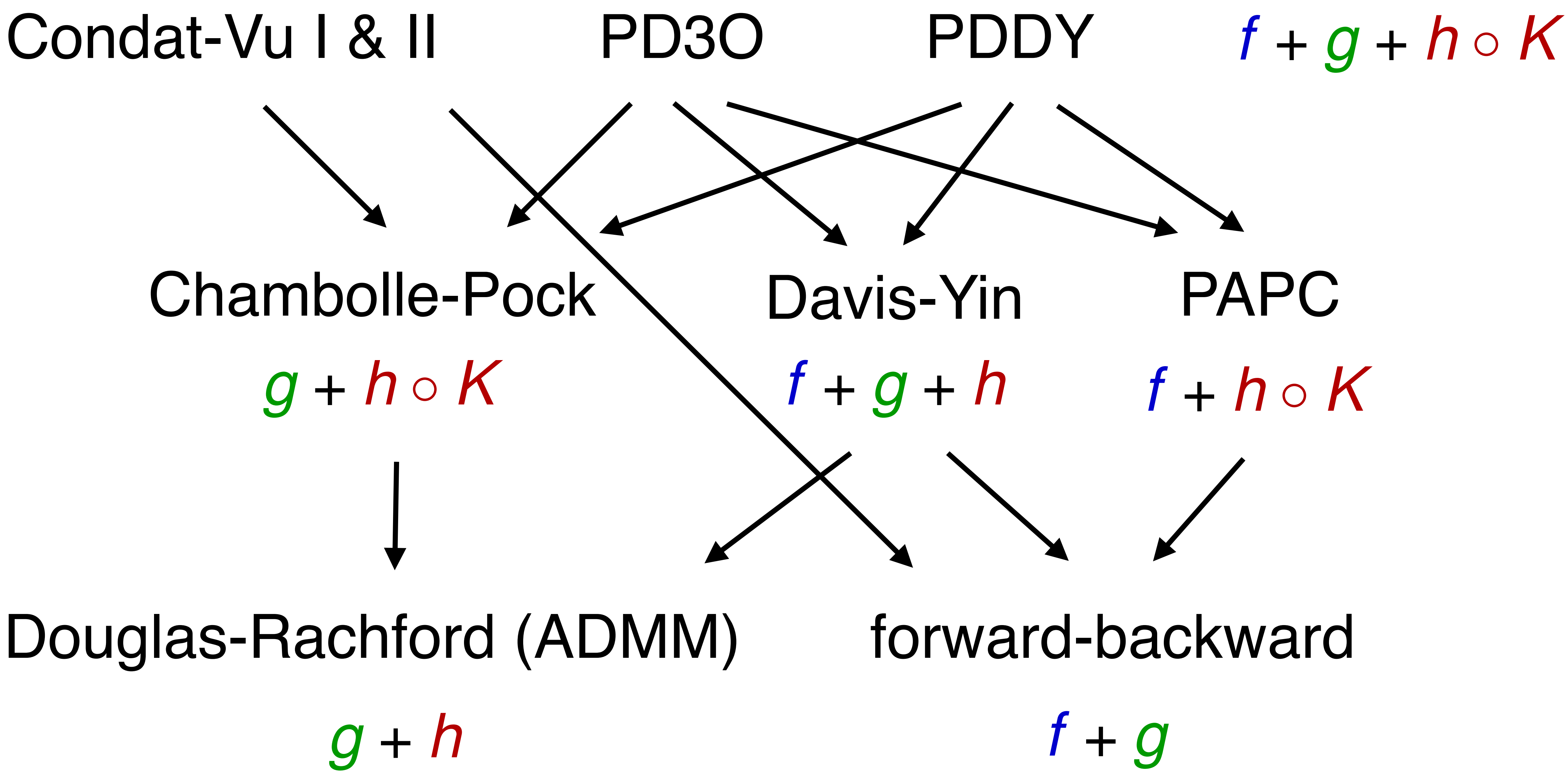
A new **randomization technique** for PDDY,
a generic primal-dual proximal splitting alg.

Note: splitting is good... if you then randomize.

- general convex case?
- acceleration?
- Bregman distances?

Bonus

Proximal splitting algorithms



4 primal-dual algorithms

minimize

$f + g + h \circ K$

Condat–Vu algorithm form I

$$\left\lfloor \begin{array}{l} x^{t+1} = \mathrm{prox}_{\gamma g}\big(x^t - \gamma \nabla f(x^t) - \gamma K^* u^t\big) \\ u^{t+1} = \mathrm{prox}_{\tau h^*}\big(u^t + \tau K(2x^{t+1} - x^t)\big) \end{array} \right.$$

Condat–Vu algorithm form II

$$\left\lfloor \begin{array}{l} u^{t+1} = \mathrm{prox}_{\tau h^*}\big(u^t + \tau K x^t\big) \\ x^{t+1} = \mathrm{prox}_{\gamma g}\big(x^t - \gamma \nabla f(x^t) - \gamma K^* (2u^{t+1} - u^t)\big) \end{array} \right.$$

PD3O algorithm

$$\left\lfloor \begin{array}{l} x^{t+1} = \mathrm{prox}_{\gamma g}\big(x^t - \gamma \nabla f(x^t) - \gamma K^* u^t\big) \\ u^{t+1} = \mathrm{prox}_{\tau h^*}\big(u^t + \tau K(2x^{t+1} - x^t - \gamma \nabla f(x^{t+1}) + \gamma \nabla f(x^t))\big) \end{array} \right.$$

PDDY algorithm

$$\left\lfloor \begin{array}{l} u^{t+1} = \mathrm{prox}_{\tau h^*}\big(u^t + \tau K x^t\big) \\ x^{t+1} = \mathrm{prox}_{\gamma g}\big(x^t - \gamma \nabla f(x^t - \gamma K^*(u^{t+1} - u^t)) - \gamma K^* (2u^{t+1} - u^t)\big) \end{array} \right.$$